CHAPTER – ONE
INTRODUCTION

1.1 Overview

Data science is an interdisciplinary field of scientific methods, processes, algorithms and systems to extract knowledge or insights from data in various forms, either structured or unstructured.

Data science is a "concept to unify statistics, data analysis, machine learning and their related methods" in order to "understand and analyse actual phenomena" with data.

It employs techniques and theories drawn from many fields within the broad areas of mathematics, statistics, information science, and computer science, in particular from the subdomains of machine learning, data mining, databases, and visualization.

Machine learning is a field of computer science that uses statistical techniques to give computer systems the ability to "learn" (i.e., progressively improve performance on a specific task) with data, without being explicitly programmed.

Machine learning tasks are typically classified into two broad categories:

- **Supervised learning:** The computer is presented with example inputs and their desired outputs, given by a "teacher", and the goal is to learn a general rule that maps inputs to outputs.
- **Unsupervised learning:** No labels are given to the learning algorithm, leaving it on its own to find structure in its input. Unsupervised learning can be a goal in itself (discovering hidden patterns in data) or a means towards an end (feature learning).

The loan predictor model come under the domain of application of supervised learning. The loan predictor model is used by the banking sector to reduce the time overhead in the decision-making capability of the organization by analysing the credentials of the applicant.

These kinds of the models have become a trending technology in the last seven years as it offers significant reduction in time and saves a lot of capital investment in the domain of document processing. These categories of mathematical models are created because of predictive modelling techniques.

Predictive modelling is a process that uses data mining and probability to forecast outcomes. Each model is made up of a number of predictors, which are variables that are likely to influence future results. Once data has been collected for relevant predictors, a statistical model is formulated. The model may employ a simple linear equation or it may be a complex neural network, mapped out by sophisticated software.

As additional data becomes available, the statistical analysis model is validated or revised.

Predictive modelling is often associated with meteorology and weather forecasting, but it has many applications in business. Bayesian spam filters, for example, use predictive modelling to identify the probability that a given message is spam. In fraud detection, predictive modelling is used to identify outliers in a data set that point toward fraudulent activity. And in customer relationship management (CRM), predictive modelling is used to target messaging to those customers who are most likely to make a purchase. Other applications include capacity planning, change management, disaster recovery, engineering, physical and digital security management and city planning.

1.2 Problem Definition

Among all industries, insurance domain has the largest use of analytics & data science methods. This data available would provide can enough taste of working on data sets from insurance companies, what challenges are faced, what strategies are used, which variables influence the outcome etc. This is a classification problem comes under the sub domain of supervised learning which comes under the machine-learning paradigm.

In this project, a predictive model has to be created by taking a data set and use that as a training set to create the model.

1.3 Objectives

Objectives are as follows:

- To gather the initial requirement specifications for the project.
- To create appropriate design diagrams for the project.
- To create software requirement specification document.
- To configure the local system according to project specification.
- To refine the data set and make it suitable to create the model.
- To create a predictive model that has a minimum of 80% accuracy.
- To create a model that can predict the whether the loan of the applicant can be approved or not.